(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 779 731 A1

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
**18.06.1997 Bulletin 1997/25**

(51) Int Cl.[6]: **H04M 3/50**, G06F 17/21, G06F 3/16

(21) Application number: **96308267.2**

(22) Date of filing: **15.11.1996**

(84) Designated Contracting States:
**DE FR GB**

(30) Priority: **15.12.1995 GB 9525719**

(71) Applicant: **Hewlett-Packard Company**
**Palo Alto, California 94304 (US)**

(72) Inventor: **Haddock, Nicholas John**
**Bristol BS1 6HJ (GB)**

(74) Representative: **Kilgannon, Denise Mary**
**Hewlett-Packard Ltd,**
**IP Section,**
**Building 2,**
**Filton Road,**
**Stoke Gifford**
**Bristol BS12 6QZ (GB)**

## (54) Speech system

(57) The present invention relates to a system for capturing and storing speech data records comprising:

means for storing parts of a speech data record in a plurality of form fields:

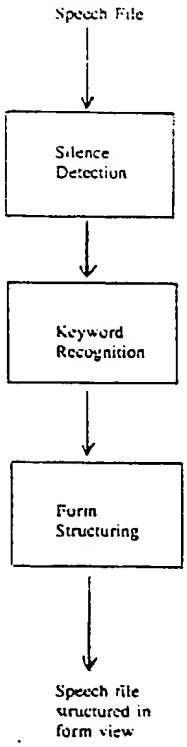
means for the user to input form field indicators:

means for recognising form field indicators:

wherein the system is operable to store speech data in a speech data record in the form

fields according to said indicators.

Fig. 2




EP 0 779 731 A1

## Description

### Technical Field

The present invention relates to devices for storing and accessing speech data.

As computing appliances shrink in size, speech will be an increasingly natural medium for entering and capturing information. The benefits of speech as an input medium are well-known. it is suitable in situations where the user is busy with their hands and eyes. and it is a quick way of capturing information. A growing range of pocket-size products allow the user to capture and store speech data in digital form and to play back voice messages.

However. one disadvantage of information held as recorded speech is that it can be arduous to review later. This invention aims to address that problem.

### Background Art

Work has been done on organising voice recordings in the form of storage folders on a computer system. Several research projects and products have demonstrated how computer-based. digital voice files can be edited using a graphical editor. Either the whole file. or sections of the file. can be moved or copied into folder areas to aid subsequent retrieval. Documented examples are Hindus. D.. Schmandt. C.. and Horner. C. 1993. Capturing. Structuring and Representing Ubiquitous Audio in ACM Transactions on Information Systems, 11 (4). October. pp. 376 - 400 and Stifelman, L. J. et al 1993. "VoiceNotes: A speech interface for a hand-held voice notetaker". Proc. InterCHI 1993. ACM, New York.

Applicant's earlier European Patent Application No. 679005 discloses a system in which a visual representation of voice data is displayed and iconic tags are used to automatically store associated parts of speech data in predefined storage areas.

Work has also been done on creating index points in voice recordings. It has been shown how index points can be stamped on voice recordings to aid subsequent retrieval of interesting sections of audio. This is common in consumer electronics products such as hi-fi cassette recorders. The paper by Degen. L.. Mander,R., and Salomon. G. 1992. "Working with Audio: Integrating Personal Tape Recorders and Desktop Computers". Proc. CHI 1992; ACM, New York discusses a system where the user can associate index markers with a voice file. A system such as described in IBM Technical Disclosure 36/09B (Sept. 1993) "Method of categorising phone messages into message logs" allows searching within the voice file for a specific keyword or phrase.

There has also been work done on eliciting voice input in a structured, form-based manner. Certain telephone answering services generate voice prompts to structure a caller's voice message (an example is described in the paper by Schmandt, C. and Arons. B., 1985. "Phone Slave: A Graphical Telecommunications Interface". Proc. Soc. Information Display. 26(1)). In effect. the caller is filling in a verbal form in response to questions such as:

"What is your name? (BEEP)": "What are you calling about? (BEEP)": etc.

Such services are becoming popular because they simplify the task of listening to message enquiries and routing them to the correct destination within a company.

### Disclosure of Invention

According to the present invention we provide a system for capturing and storing speech data records comprising:

means for storing parts of a speech data record in a plurality of form fields:

means for the user to input form field indicators:

means for recognising form field indicators:

wherein the system is operable to store speech data in a speech data record in the form fields according to said indicators

By providing for form field indicators to be input by the user. the invention enables the indexing of speech data as it is recorded and permits an interaction technique which allows structure and some content to be extracted from a voice record. thus making it easier to review the recording later and to integrate it with other data.

Preferably the system further comprises:

a plurality of storage areas for storing speech data records:

means for inputting storage area indicators:

means for recognising storage area indicators:

wherein the system is operable to store speech data records in the storage areas according to said indicators.

In this way the speech data can be divided into categories convenient for the user eg. phone numbers. to-do items etc.. as well as structuring the individual speech records.

In an embodiment to be described, the indicators are keywords spoken by the user and the system comprises memory means for storing a set of key words and means for recognising a key word when spoken.

Optionally, the system may comprise means for detecting a keyword marker in speech data; means for triggering key word recognition on speech data associated

with a keyword marker: and means for storing speech data according to the indentity of the associated key word.

A keyword marker may be a pause of predetermined duration in the speech data or may be generated by the user operating a predefined input device. such as a button.

## Brief Description of Drawings

Preferred embodiments of the present invention will now be described. by way of example. with reference to the accompanying drawings:

Figure 1 shows a handheld computer implementing the present invention:

Figure 2 indicates the main system components required for implementing the present invention.

## Best Mode for Carrying Out the Invention. & Industrial Applicability

Figure 1 shows a handheld computer 10 comprising a display screen 12 and a set of keys 14 for user input. The computer contains four data areas. or 'applications': a phone book. diary. to-do list. and messages. Each application is a list of entries. and each entry in the list is a form with several fields. Optionally. text can be entered into each field for display to the user. If the system has automatic word recognition capability. text may be entered automatically by the system.

The screen 12 shows a form-based phone book entry 16. The entry 16 includes six fields: Name: Home Number: Business Number: Fax Number: Address: Comment. An audio icon 18 indicates that a field contains speech. Speech can be played back in clips from its field location. or as the original whole record.

New voice recordings are by default added to a 'general' list. which displays simple header information about the items. Further to this. if the voice record begins with the name of a recognised application, such as "Phone book entry", then the voice record is filed into that area: if not, it remains in the general list. If an application name is recognised. then the rest of the voice record is searched for any keyword labels corresponding to form fields. Key word (or phrase) labels are assumed to be preceded by a pause. The speech content following the label. up to the next recognised field name. is then associated with this field in the form.

In order to organise incoming speech data in this way. the system starts by recognising the first section of speech as the name of one of the applications: messages. phone book entry, todo item, or diary entry. This determines which keyword labels (for field names) are then looked for in the remainder of the speech file.

In order to detect keyword labels and build a strcuture from them, three technical components are involved in series. as indicated in Figure 2:

1. Silence detection
2. Keyword recognition
3. Form structuring

1. Silence detection - once the speech file has been recorded (and placed in the 'general' list). it is scanned for any pauses longer than one second. Silence detection is a standard speech processing technique. and can be implemented in a number of ways. The energy level is measured throughout the voice file, and speech is assumed to be present whenever this level exceeds a threshold. The threshold itself is set in an adaptive manner. depending on the background noise present. The paper by O'Shaughnessy. D 1987. Speech Communication. New York: Addison-Wesley describes one way of implementing silence detection.

2. Keyword recognition - a standard class of speech recognition technology is used. operating within a small vocabulary (just the field names): the recogniser is for continuous speech. speaker-independent recognition based on sub-word Hidden Markov Models. An available product is by Young. S. J.. Woodland. P.C. and Byrne. W. J. HTK: 'Hidden Markov Model Toolkit 1.5'. 1993 of Entropic Research Laboratories. Inc. and a published paper is: Young, S. J. and Woodland. P.C. 1993. "The HTK Tied-State Continuous Speech Recogniser" Proc. Eurospeech '93.

The pauses detected at the silence detection stage are the initial anchor points for keyword label recognition. At each anchor point. an attempt is made to match the initial section of subsequent speech against one of the stored keywords or phrases. For example. within the phone book application. after each pause the system is looking for one of:

NAME IS
HOME NUMBER IS
BUSINESS NUMBER IS
FAX NUMBER IS
ADDRESS IS
COMMENT IS

In addition, each keyword/phrase may be followed by 'garbage' phonemes. Garbage is defined as any sequence of phonemes other than the relevant keywords. This is because each keyword label will be followed by material such as "(Name is) Janice Stevens". and the system is not attempting to recognise the name Janice Stevens.

The user may have unintentionally paused during recording. and in these cases the pause will not necessarily be followed by a keyword label. The garbage model also detects the non-keywords which may follow these natural juncture pauses within the speech. Hence at each recognition stage the recogniser is looking for one of the recognised keyword labels, OR garbage.

3. Form structuring - the final component must create the form structure from the segments defined in the silience detection stage. some of which have been labelled with keywords by in the next stage. A segment which begins with a recognised keyword label is associated with the corresponding form field. In addition. all subsequent speech segments. up to (and not including) the next speech segment which starts with a keyword label. are associated (in sequence) with this form field. If a given keyword label occurs at the beginning of more than one segment. then the second occurrence takes precedence. and the former occurrence is ignored.

For example. if while driving along. a user is overtaken by a truck laden with useful looking contact information. they could quickly record the following voice memo using a system according to the invention:
"Phone book entry ... Business number is 408 927 6353 ... Name is Hamilton & Co Removals ... Comment is the truck says something about extra heavy items being a speciality.
could get that piano shifted at last".

Given the spoken keywords (underlined). the recorded note will be added as a new entry in an electronic phone book. and the speech will be segmented into three of the different fields in the phonebook form shown on the display in Figure 1.

The present inventtion provides an interaction technique which allows structure and some content to be extracted from a voice record. thus making it easier to review the recording and integrate it with other data. In particular. it introduces a technique for automatically extracting form structure from a voice recording. This is accomplished by allowing the user to insert indicators (such as keywords) into their speech. as the speech is being recorded, to form index points.

If the recognition capability of the device is sufficiently good. there may be no need for keyword markers (pauses in the above embodiment) at all.

In an alternative embodiment. button presses could be used as keyword markers instead of (or in combination with) silence detection. Here the user would press a button to indicate when a keyword was about to be uttered.

Another alternative to the embodiments described above involves designing the system so that the form field indicators were not necessarily immediately preceding the speech data to which they correspond. For example. the form field indicator could follow the relevant speech data or could be surrounded by it.

## Claims

1. A system for capturing and storing speech data records comprising:

   means for storing parts of a speech data record in a plurality of form fields:

   means for the user to input form field indicators:

   means for recognising form field indicators:

   wherein the system is operable to store speech data in a speech data record in the form fields according to said indicators.

2. A system according to claim 1 further comprising:

   a plurality of storage areas for storing speech data records:

   means for inputting storage area indicators:

   means for recognising storage area indicators:

   wherein the system is operable to store speech data records in the storage areas according to said indicators.

3. A system according to claim 1 or claim 2 wherein the indicators are keywords spoken by the user and the system comprises memory means for storing a set of key words and means for recognising a key word when spoken.

4. A system according to claim 3 comprising means for detecting a keyword marker in speech data:

   means for triggering key word recognition on speech data associated with a keyword marker:

   means for storing speech data according to the indentity of the associated key word.

5. A system according to claim 4 wherein a keyword marker is a pause of predetermined duration in the speech data.

6. A system according to claim 4 wherein a keyword marker is generated by the user operating a predefined input device.

Figure 1

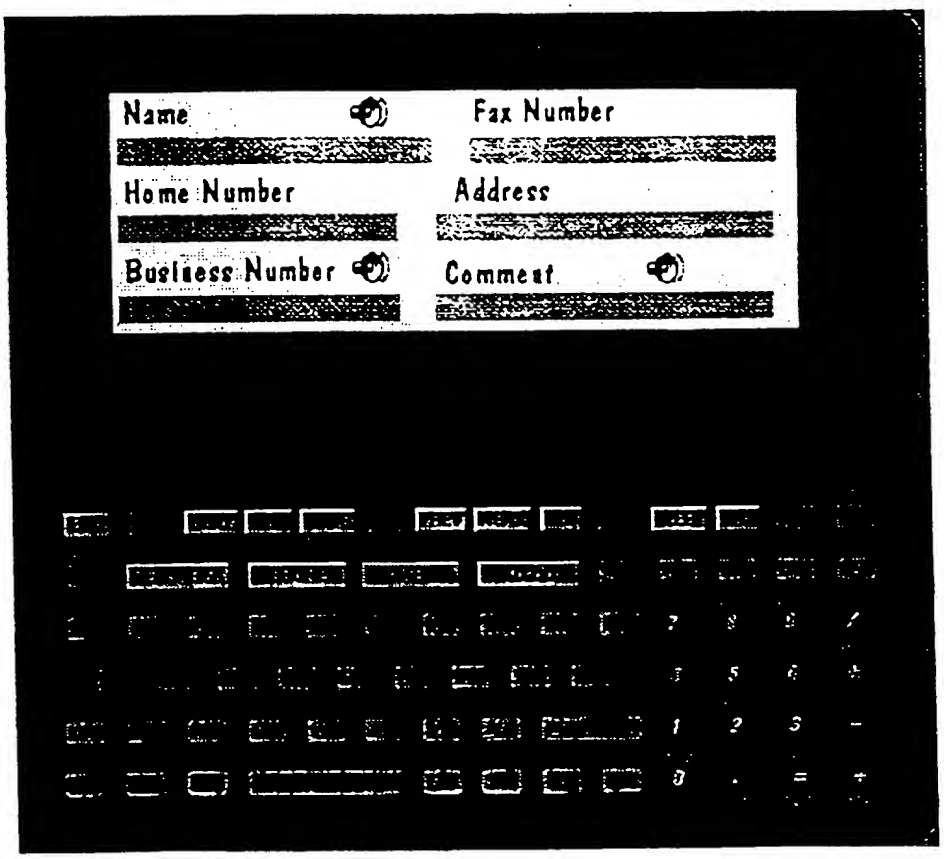
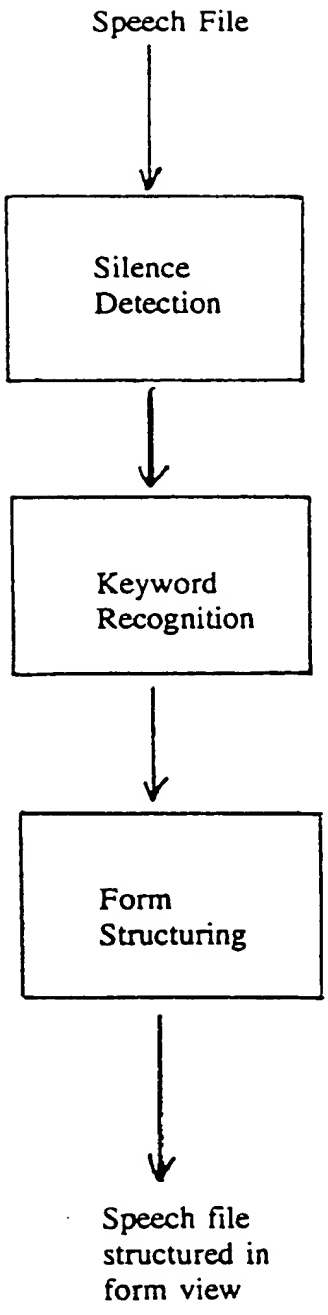



Figure 1

Fig. 2


Speech File
Silence Detection
Keyword Recognition
Form Structuring
Speech file structured in form view

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 96 30 8267

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| X | UIST 95- PROCEEDINGS OF THE ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY-, 14 - 17 November 1995, PITTSBURGH (US), pages 171-179, XP000602486 LISA J. STIFELMAN: "A TOOL TO SUPPORT SPEECH AND NON-SPEECH AUDIO FEEDBACK GENERATION IN AUDIO INTERFACES" * page 171, right-hand column, line 33 - page 172, left-hand column, line 14 * * page 174, left-hand column, line 9 - line 16 * | 1-6 | H04M3/50 G06F17/21 G06F3/16 |
| A | RESEARCH DISCLOSURE, no. 361, May 1994, HAVANT(GB), page 234 XP000453941 "AUTOMATIC QUICK SEARCH SYSTEM IN DICTATION APPARATUSES" | | |
| A | WO 93 07562 A (RIVERRUN TECHNOL) | | |
| A | IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 6B, June 1993, NEW YORK US, pages 169-170, XP000377340 "CATEGORICAL STORAGE OF VOICE MAIL MESSAGES" | | TECHNICAL FIELDS SEARCHED (Int.Cl.6) H04M G06F |
| A | IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 8, August 1993, NEW YORK US, pages 405-407, XP000390273 "METHOD FOR ENHANCED MESSAGING SERVICE" | | |
| A | IBM TECHNICAL DISCLOSURE BULLETIN, vol. 38, no. 6, June 1995, NEW YORK US, page 635 XP000520801 "BATCH PROCESSING OF AUDIO MESSAGES" | | |

-/--

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 11 March 1997 | Vandevenne, M |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 96 30 8267

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| A | IBM TECHNICAL DISCLOSURE BULLETIN, vol. 37, no. 1, January 1994, NEW YORK US, pages 559-560, XP000428882 "METHOD OF AND APPARATUS FOR CATEGORIZING PHONE MESSAGES" --- | | |
| A | WO 92 02009 A (YOUNGER) ----- | | |
| | | | TECHNICAL FIELDS SEARCHED (Int.Cl.6) |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 11 March 1997 | Vandevenne, M |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)